# Only Time Can Tell: Discovering Temporal Data for Temporal Modeling

Laura Sevilla-Lara
University of Edinburgh

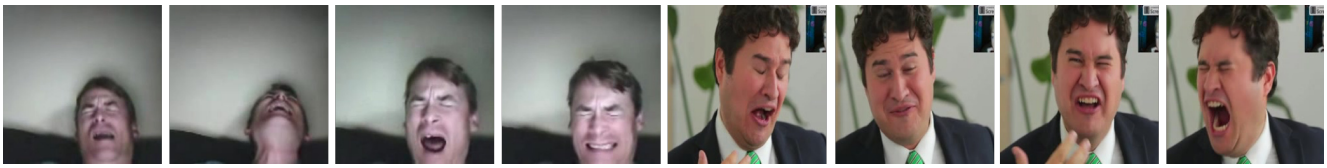
Shengxin Zha
Facebook AI

Zhicheng Yan
Facebook AI

Vedanuj Goswami
Facebook AI

Matt Feiszli
Facebook AI

Lorenzo Torresani
Facebook AI



**Can you guess these actions? "yawning", "sneezing" or "crying"?** Temporal information is essential to discriminate some actions, while for others it is redundant. Shuffling frames in time removes temporal information, revealing the actions where it actually matters. (Solution at the end of the paper.)

## Abstract

*Understanding temporal information and how the visual world changes over time, is a fundamental ability of intelligent systems. In video understanding, temporal information is at the core of many current challenges, including compression, efficient inference, motion estimation or summarization. However, in current video datasets it has been observed that action classes can often be recognized without any temporal information, from a single frame of video. As a result, both benchmarking and training in these datasets may give an unintentional advantage to models with strong image understanding capabilities, as opposed to those with strong temporal understanding. In other words, current datasets may not reward good temporal understanding, potentially hindering progress. In this paper we address this problem head on by identifying action classes where temporal information is actually necessary to recognize them and call these "temporal classes". Selecting temporal classes using a computational method would bias the process. Instead, we propose a methodology based on a simple and effective human annotation experiment. We remove just the temporal information, by shuffling frames in time, and measure if the action can still be recognized. Classes that cannot be recognized when frames are not in order, are included in the temporal set. We observe that this set is statistically different from other static classes, and that performance in it correlates with a network's ability to capture temporal information. Thus we use it as a benchmark on current popular networks, which reveals a series of interesting facts, like inflated convolutions bias networks towards classes where motion is not important. We also explore the effect of training on the temporal set, and observe that this leads to better generalization in unseen classes, demonstrating the need for more temporal data. We hope that the proposed dataset of temporal categories will help guide future research in temporal modeling for better video understanding.*

## 1. Introduction

Temporal information has been shown to be important for recognition. In the biological setting, Johansson's point-light experiments [18] show that moving dots alone contain enough information to understand actions, when the points-lights are placed at the joints. Similarly, Heider and Simmel [11] find that simple figures (like squares or triangles) in motion can convey rich story information, including intention and emotions like anger, joy. Fig. 1(a) shows two frames from these experiments; neither the action nor a story can be recognized easily in the absence of temporal information.

In computational settings, temporal features (optical flow, spatio-temporal convolutions, or both) have been shown to be useful for video understanding. For some datasets and models, using optical flow as input actually

yields higher accuracy than using RGB images [32, 3, 36]. However, a single image is often informative enough to predict the label with good confidence. This leads to the question *when do we need temporal information to understand an action?* Fig. 1(b) shows some examples from Kinetics [20], where the action class is easy to guess. Furthermore, in recent two-stream models such as I3D [3] fusing the predictions of the stream operating on optical flow (the temporal stream) to that operating on RGB frames (the spatial stream) produces rather small gains. Do these observations mean that given the image content, motion information is redundant?

We argue that actually these observations do not reflect the importance of motion in video understanding, but rather the distribution of categories in some of the video datasets. Based on this hypothesis, we propose to examine some of the most widely used datasets, and identify classes where temporal information is necessary to recognize them. We call these "temporal classes." Our analysis is based on human perception rather than on the performance of computational models as we do not want to tie the definition of temporal categories to a specific model's performance. In particular, given a collection of videos , we compare human performance on a video with shuffled frames and on the same video with the original ordering of frames. We use this difference as a cue for how necessary temporal information is to recognize a class. We then select the classes where this difference is largest. We use these categories as a benchmark to measure how well different architectures perform. Our analyses provides interesting observations such as that the ranking of models on the temporal classes is different compared to that on the full datasets and that temporal models trained on temporal classes have stronger generalization performance. In summary, we make the following contributions:

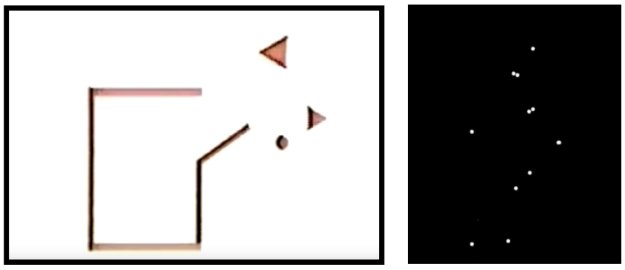
- A *methodology to discover the importance of temporal information* in action classes based on human perception. This allows us to untie the definition of temporal relevance from the performance of existing computational methods.

- The *temporal set of classes, resulting from using our proposed methodology on current video classification datasets*. We identify the set of categories containing more temporal information and what they have in common, and examine their computational properties. We will release all of the collected human judgments for further analysis by the community.

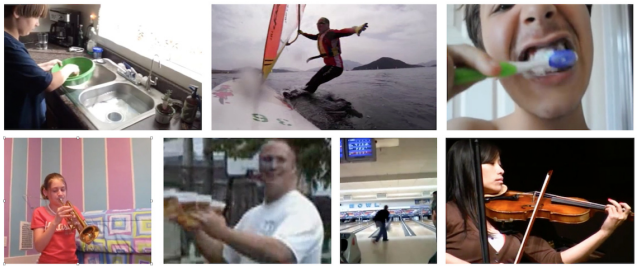
- We *benchmark current video models*, and observe that the ranking changes with the inclusion or omission of temporal categories, revealing hidden biases and strengths. This also suggests that the distribution of the videos in the temporal categories actually has different structure.

- We use the *temporal categories for training temporal models*. We observe that training on the temporal set produces stronger temporal features. We also observe that training on this dataset leads to a better ability to generalize in the transfer learning setting of training on "seen" categories and testing on "unseen" categories.

We hope that our proposed "meta-dataset" of temporal categories will help guide future research in temporal modeling for better video understanding.



(a) Temporal task.



(b) Static task.

Figure 1. **The relevance of temporal information varies vastly by task.** In one end of the spectrum (a) shows sample frames from Heider and Simmel [11] (left) and Johansson [18] (right). Images alone do not give much information, but when the video is played, an action or a even a story comes to life. These experiments highlight the richness of temporal information in visual understanding. At the other end (b) shows sample frames from Kinetics. It is easy to map images to *playing trumpet, brushing teeth, washing dishes, drinking beer, playing violin, windsurfing, bowling*. In this task, temporal information is completely redundant.

## 2. Related Work

**Temporal Modeling**. Over the last years video understanding has made rapid progress. While modeling spatial information on videos has largely been addressed leveraging advancements in image understanding methods, modeling temporal information has sprouted new and creative techniques. These techniques can be broadly categorized in three families: two-stream, temporal convolutions, and recurrent methods. The first, pioneered by

Simonyan and Zisserman [32], consists on having two parallel networks: the spatial stream aimed at modeling image content, which takes RGB as input, and a temporal stream aimed at modeling motion content, which takes a small number of neighboring flow fields as input. The predictions of the two streams are combined at the end. Different variants of this general framework arise using different architectures for the streams [37, 8]. These techniques are described as frame-based, since they do not explicitly model inter-frame information. Temporal convolution methods [35, 36, 38, 25, 14] extend the success of 2D convolutions in image understanding to include convolutions across frames that are adjacent in time. Recently, TSM [24] augments 2D CNN with temporal shift modules to achieve temporal modeling. Finally, recurrent convolutional networks [34, 39, 6] and relational networks [41] have been used for learning video representations across multiple clips, most often by applying a 2D CNN as a feature extractor on the individual frames/segments and using a recurrent or relational network on the CNN features to model temporal information. Some of the latest methods combine the two-stream with the temporal convolution approach, like I3D [3]. In this paper we will focus on this hybrid family of methods, that include two streams and temporal convolutions, since they will allow us to experiment and compare the impact of testing and training on temporal data when using flow vs using images, and using frame-based models vs using temporal convolutions.

**Video Datasets**. Progress in video understanding has been greatly propelled by the growth of video datasets in data scale and richness of taxonomy. Two of the first datasets in this area were UCF-101 [33], HMDB-51 [22], which have been widely studied since the early 2010's. These are single-label videos of human actions in 101 classes and 51 classes, respectively. The first large-scale dataset to appear was Sports-1M [19], which provides 1M videos of sport actions in 487 classes. Although the labels are predicted by tagging algorithms and thus noisy, Sports-1M has consistently improved the quality of models pre-trained on it. In the last few years, researchers have pushed video classification to more challenging scenarios by using human annotation to collect large-scale datasets with different characteristics. For example, the Something-Something [12] dataset contains categories of generic actions independent of objects, with the goal of obtaining models that reason over physical aspects of actions and scenes as opposed to high-level action concepts. Kinetics [20] includes 400 categories of human actions in different environments, combining the success case of UCF101 with large-scale. Moments-In-Time dataset [27] contains 1M short videos corresponding to over 300 event classes. Unlike past datasets that focus on third person videos, Charades [31] focuses on egocentric first-person videos that happen around the house. These datasets are often created by first deciding on a taxonomy of classes and then looking for videos containing instances of such classes, or recording them [27]. This process may bias the distribution of videos based on the search engine, and thus SOA [28] proposes a dataset where the videos are chosen first, and then labeled using the taxonomy of classes that naturally arises. While this is a more resource intensive labeling process, it yields videos often with multiple labels, and datasets with less biased label taxonomy. Other recent datasets also extend the types of video tasks to action localization(*e.g*. THUMOS 14 [17], ActivityNet [2], AVA [13], HACS [40]), video object detection (*e.g*. VLOG [10]), action sequence classification [23], and video prediction (*e.g*. Epic-kitchens [5]). Despite the tremendous efforts of dataset collection, it is not well studied yet what action classes substantially require temporal modeling to recognize as opposed to still appearance modeling.

**Temporal Ordering and Frame Shuffling**. Temporal ordering in the video is arguably important for video understanding. A rich set of methods have explicitly exploited it for video recognition, including ranking pooling [9, 4], and RNN-based methods [7, 34]. Furthermore, Misra et al. [26] use it as a self-supervised task to learn better temporal models, where a network is trained to predict the original ordering of a shuffled set of frames. Other recent work has explored the effect of shuffling to measure the robustness of a network to temporal disarray [30, 16], observing that even networks using temporal convolutions can be surprisingly insensitive to shuffling frames within clips. Finally, perceptual studies also have investigated the effect of shuffling in space and time [21]. In contrast, in this work, we conduct a human perception study, and use frame shuffling to identify action classes where the action recognition performance of human is substantially compromised by the randomly shuffled temporal ordering.

## 3. Discovering Temporal Classes

In this section we describe the process to discover the categories for which temporal information is important, which will be part of the temporal set of classes. We first describe the perceptual test in detail and then discuss our findings.

### 3.1. Perceptual Test to Discover Temporal Classes

**Motivation.** How can we measure if temporal information is necessary to identify an action in a video? We could, for example, compare the accuracy of a method using the entire video as input to the accuracy using a single frame. However, going from the full video to a single frame eliminates not only the temporal information, but also some

appearance information, since we would be excluding most of the frames. Further, recent studies have shown [16] that some frames make recognition easier than others, raising a more complicated "frame selection problem". Instead we maintain all image information and remove all temporal information by shuffling frames in time, using the resulting video as input. We could choose an action recognition network and observe the performance drop with and without temporal shuffling. However, this approach would tie the selection of temporal classes to a specific computational model. This is risky for two reasons. First, it conflates the inherent complexity of identifying a class and the temporal information. Second, we do not know the architecture's ability to represent temporal information in the first place. Our intent here is not to find classes where a specific model does poorly, but to define a "meta-dataset" of temporal classes that will spur innovation in temporal modeling. Based on these arguments we propose to show the shuffled videos to human subjects and define temporal relevance of a class based on the drop in accuracy in human recognition performance when the video frames are shuffled in time.

**Stimuli.** We empirically found that watching videos where frames are shuffled in time can be discomforting. It is hard for humans to attend to so much change in the scene. Even if we slow down the frame rate there is a fundamental problem: we are simply replacing the true temporal information with random temporal information; the human perceptual system will still attempt to connect the frames in time. For this reason we render the video on a canvas containing space for two frames side by side. At any time we black out one half (either the left or the right half) of the canvas and render the current frame in the other half. The next frame is then shown in the half that was previously blacked out. This left-right alternation eliminates any perception of motion (real or random), producing a slide-show effect where image information is retained but temporal information is removed. Video examples are included in the supplementary material.

**Task.** Action recognition networks are typically trained for multi-class classification, where given an input video, they classify the action in it. Ideally we would like to expose human annotators to the exact same task. However, this is difficult in practice, since the number of classes is in the order of hundreds. Annotators would have a hard time remembering all the classes in the taxonomy. Another option is to ask the binary question "is X action happening in this video?". The problem there is that we only have positive examples, and generating compelling negative examples is not trivial. Instead, for each video we show as possible answers only classes within a semantic group. For example, if a video has the label "barbecueing", then we will show as possible answers other cooking categories, which includes classes like "making a sandwich", since they are more likely to be visually similar. [1] This restricts the options that we give to the annotator, from the order of hundreds to 5-15 options. While this may be considered making the task easier, in practice it would be rare that categories across groups would be accidentally mistaken for each other.

**Datasets.** We discover temporal classes from the most relevant action classification datasets. We use Kinetics [20] (400 classes, 290K videos), for being the most widely used set. We also use Something-Something [12] (174 classes, 100K videos) because it was specifically designed to be action-centered and independent of the objects appearing in the scene which makes it particularly relevant to our study.

**Selection of Temporal Classes.** We show between 15 and 30 videos per class to a set of 20 different human annotators. For each class $i$ we compute the average accuracy of the annotators when they see the videos shuffled $\text{Acc}_s(i)$ and for the control $\text{Acc}_c(i)$, when the videos are not shuffled. We select the groups where the control accuracy is larger than $70\%$ (*e.g.*, the classes names are sufficiently clear) and the drop in accuracy $\text{Acc}_c(i) - \text{Acc}_s(i)$ is larger than a threshold ($40\%$ in Kinetics, and $60\%$ in Something-something, since performance drops have different statistics). We also include those action classes that have been the source of confusion more than $20\%$ of the times, since they are necessary to actually observe the confusion (*e.g.* plugging and unplugging, opening and closing, etc).

### 3.2. What Actions Actually Require Temporal Information?

In total, we discover 50 temporal classes, where 32 come from the Kinetics datasets and 18 from Something-something. We call this subset the temporal set of classes. The total amount of videos is 35,045, where 32,081 are for training and 2,964 for testing.

In Table 1, we show the actions where accuracy drops most when we remove temporal information, by group and by class, respectively. As expected, temporal information is more important when the scene or objects in are not discriminative for the action recognition task, for example in videos showing different forms of dancing, or different skiing styles. These are classes where the background, clothing, and objects tend to be fairly constant across categories or uninformative for action recognition. We also observe that facial actions are particularly sensitive to frame shuffling, making actions like yawning and sneezing difficult to

[1] These semantic groups are provided in the Kinetics dataset, but for the Something-something dataset we design them manually, using the verb as guidance.

| Class | Confused with | ∇Acc (%) |
|---|---|---|
| Ski jump. | Ski, Ski slal. | 73.33 |
| Play cards | Shuffle cards | 66.66 |
| Ski crossc. | Ski jump., Ski slal. | 60.00 |
| Drop kick | Wrestle, High kick | 60.00 |
| Dance macar. | Breakdance, Tap | 60.00 |
| Shake head | Sneeze, Headbang | 53.33 |
| Sneeze | Yawn, Blow nose | 46.66 |

Table 1. **What is motion for?** Classes from the Kinetics dataset where human recognition accuracy drops most when temporal information is removed.

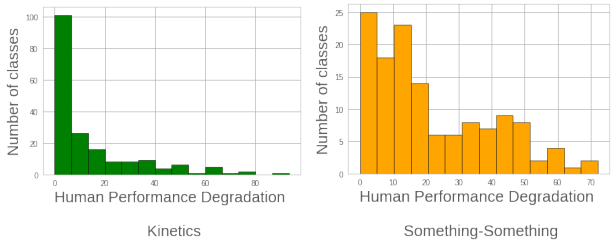



Figure 2. **Histogram of performance drop (%).** Something-Something contains more categories where temporal information is necessary.

tell apart without temporal information.

Fig. 2 shows the histogram of performance drop of all classes after temporal shuffling. The proportion of classes that require temporal analysis is small for both datasets. Something-Something shows a larger proportion of categories requiring temporal information for discrimination. This is not surprising since this dataset was designed precisely to avoid strong correlations between object categories and action classes. Perhaps what may be more surprising is that still many of the classes in such dataset can be correctly identified from shuffled frames. This highlights the need for the proposed methodology: it is difficult to design a dataset of actions that require strong temporal modeling for accurate recognition, while it is much easier to discover temporal classes using our simple perceptual test.

# 4. Analysis of the Temporal Classes

## 4.1. Experimental Details

**Datasets.** We use two sets of classes in our experiments: *Temporal-50*, which is the set of 50 temporal action classes where performance decreases most. And *Static-50*, the 50 classes where human accuracy decreases least. Both sets are measured in Sec. 3.2. The number of videos in this set is comparable to that of Temporal-50. While there is a larger portion of classes where accuracy decreases a negligible amount, we choose 50 at random to keep the number of classes and videos comparable to Temporal-50. Note that both sets contain the same proportion of classes from Kinetics, and Something-something, to factor out the dataset bias component.

**Architectures.** We choose a suite of architectures that are top-performing and most widely used. They include some that use temporal information (*e.g.* through optical flow or temporal convolution) and others that do not. In particular, we use: R2D which is a frame-based model, that uses the ResNet [15] architecture with 18 layers, and it is trained with a cross-entropy loss; R3D [35] which is similar to R2D, but instead of processing each frame separately it includes convolutions in the temporal dimension; I3D [3] which is also a 3D architecture, that extends inception to use convolutions in the temporal dimension; and R(2+1)D which is a recent architecture [36] that achieves state-of-the-art accuracy. The main difference compared to R3D is that R(2+1)D factors the 3D convolutions into 2D filters in space and 1D filters in time. This separability helps the optimization, making this the strongest of the models we use.

**Training and Evaluation Details.** Unless otherwise specified, all training experiments are performed with the same parameters: 75 epochs, from which 10 are warm-ups and the remaining 65 follow a cosine function, where the base learning rate is 0.0025 and gamma is 0.0001, on 4 GPUs at a time. The batch size is 16. We use multiple crops of size 112×112. All models are trained from scratch. Unless otherwise specified, the reported evaluation of models is done using a clip size of 16 frames. We report only the video-level top-1 accuracy in the main paper for brevity.

## 4.2. Computational Properties of Temporal Classes

In this section we perform a series of computational sanity checks, that help validate the nature of the temporal and static classes. This is, that temporal classes are actually those where temporal information matters, and the static classes those where temporal information is redundant. We argue that a randomly chosen set of classes is very unlikely to satisfy all our sanity checks.

**Temporal and static classes are statistically different.** The first sanity check is to make sure that the behavior of networks is different in the temporal and static classes in a statistically significant manner. Given a network $N$, we compute the per-class accuracy of the network $\text{Acc}_N(i)$ of all classes $i$. We now compute all accuracies of the classes in the Temporal Dataset $T$, as $A = \{\text{Acc}_N(i) \mid i \in T\}$. We also compute the accuracies in the Static Dataset $S$, which is the set $B = \{\text{Acc}_N(i) \mid i \in S\}$. We use the Kolmogorov-Smirnov test [1] to measure if $A$ and $B$ are sampled from the same distribution, for multiple networks $N$ (R(2+1)D and R2D). We observe that the null hypothesis

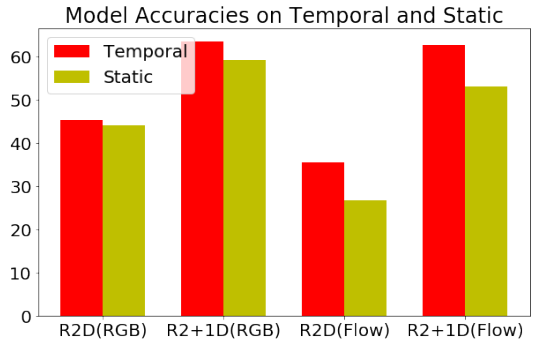
can be rejected with probability $p < 0.1$, suggesting that the behavior of the networks on the datasets $T$ and $S$ are sampled from different distributions. As a control experiment, we randomly sample two sets $T'$ and $S'$, and observe that $p > 0.4$. We conclude that the behavior of networks is statistically different in the temporal and static classes.

**A network's ability to capture temporal information is correlated with its accuracy in temporal classes.** Recall that one of our goals with the discovery of temporal classes is to use it as a benchmark to measure a network's ability to capture temporal information. In this sanity check we use pairs of networks where we do know their relative abilities to capture temporal information. In particular we use the pair R2D and R(2+1)D, which are identical except for the fact that R2D is frame-based and R(2+1)D includes temporal convolutions. Thus it is safe to say that R(2+1)D has a stronger ability to represent temporal information than R2D. We also use the pair RGB and flow. If temporal classes actually contain videos where temporal information is necessary, we should observe that R(2+1)D has an advantage over R2D on temporal classes, and so does using flow as input over using RGB. Of course, R(2+1)D has larger capacity, and it will typically always outperform R2D. We use the accuracy on static classes to control for the capacity of the networks, and measure *the difference between the accuracy in temporal and Static classes*.

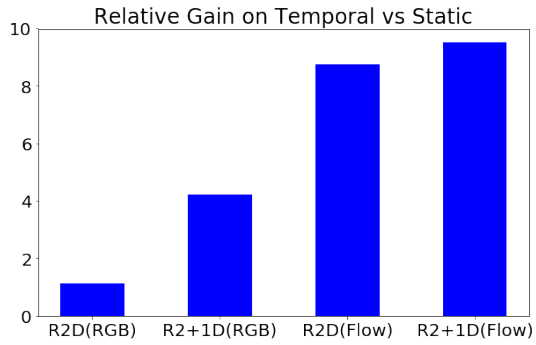
Results are shown in Fig. 3. The accuracy of the R2D network using RGB as input is similar in temporal and static classes (Fig. 3(a)). However, when we use the R(2+1)D architecture, which adds temporal convolutions, the network does better on temporal than static classes. This is consistent with our hypothesis, that networks that model temporal information better perform better in the temporal classes. Similarly, we observe better performance on temporal classes when using flow as input, than when using RGB as input. This difference is more clearly observed in Fig. 3(b), which shows the correlation between the networks' ability to capture temporal information, and its gain in temporal classes, as expected. At the same time, we do not observe this pattern in the control experiment of sampling subsets of classes at random. With this we conclude that *the relative gain in temporal classes compared to the static classes is a good measure for evaluating a network's temporal capabilities*.

## 5. Temporal Classes as a Benchmark

In this section we evaluate the ability to capture temporal information of a set of current state-of-the-art networks. We propose to use two measurements, based on the experiments of Sec. 4.2. The first measurement is the accuracy on temporal classes. The second, is the relative gain in ac-




(a) Absolute Accuracy (%).




(b) Relative Accuracy (%).

Figure 3. **The better a network represents temporal information, the better it performs on temporal classes.** R2D(RGB) has similar performance in the temporal and static classes (a). However, R(2+1)D(RGB) which adds temporal convolutions and shows better performance on temporal than static classes. This is consistent with our hypothesis, that temporal classes help identify a network's ability to model temporal information. This pattern is consistent when we switch from RGB to Flow. This difference is clearly observed in (b).

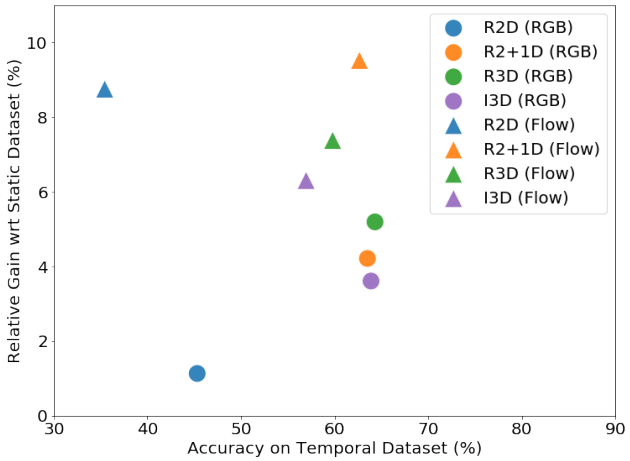



Figure 4. **Relative Gain vs. Accuracy in Temporal Dataset.** This plot gives a more complete picture, since it contains both relative and absolute accuracies.

| Network | Input | Relative Gain (Δ%) | Traditional Accuracy (%) |
|---|---|---|---|
| R2D | RGB | 1.13 | 41.14 |
| R2D | Flow | 8.74 | 26.87 |
| R(2+1)D | RGB | 4.21 | 56.01 |
| R(2+1)D | Flow | **9.50** | 51.53 |
| I3D | RGB | 3.60 | **57.37** |
| I3D | Flow | 6.29 | 49.00 |
| R3D | RGB | 5.19 | 56.40 |
| R3D | Flow | 7.36 | 50.24 |

Table 2. **Evaluation of popular action recognition networks.** The ranking of methods changes when we use the traditional score and when we use the proposed temporal score.

curacy on temporal classes with respect to the accuracy on static Dataset, as we have observed that it correlates with a networks ability to represent temporal information.

We argue that these two measurements are complementary. While the first captures the actual ranking on the dataset, this metric is influenced for example by the capacity of the network, *i.e*. a network may perform better in this dataset simply because it is larger, but not necessarily because the architecture represents temporal information better. We try to account for this by using a second measurement, which is the relative gain in accuracy on temporal classes with respect to the accuracy on static classes. We compute the relative gain simply as the difference between the accuracy on temporal classes and static classes. We compare to the traditional accuracy, computed as the average per-class accuracy over all classes. Results of evaluating a suite of popular methods are shown in Table 2.

We make several observations. First, using optical flow shows to be more useful. This reveals that the perception that optical flow is not useful in recent datasets is more a product of the task (i.e., solving Kinetics) than the true nature of the visual world. Second, we observe that the top-performing method (I3D) according to traditional accuracy, scores poorly in the temporal score. This is possibly a direct consequence of the training process using inflation based on image filters learned from ImageNet. While this process is useful for action recognition, it biases the network to excel in static classes. Since R(2+1)D and R3D are not pre-trained on Imagenet, their overall performance on the datasets is lower, but they do shine on the temporal classes. In Fig. 4 we show results of the two measurements together: accuracy in temporal classes and relative gain. This figure allows us to have a more complete picture, where both the absolute and relative accuracies are represented.

Finally, for further analysis, in Fig. 5 we plot the accuracy on temporal and static classes with respect to the traditional accuracy. We observe that the accuracy on static classes is rather correlated with the traditional accuracy, showing that overall current datasets are dominated by static classes. However, accuracy on temporal classes does not follow such a strong correlation, and actually shows a "plateau" effect on the accuracy on temporal classes. Note that the difference in accuracy among models with flow input is quite large, showing a strong ranking of about 3% difference among R(2+1)D, R3D and I3D.

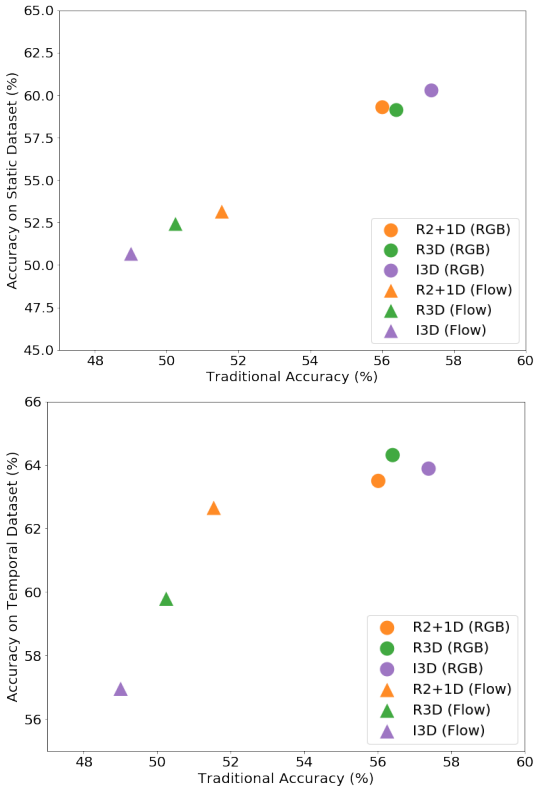



Figure 5. **Accuracy on Static and Temporal Classes vs. Traditional Accuracy.** The top figure shows the accuracy of several networks on the static classes with respect to the traditional accuracy. It is obvious that these two measurements are very well correlated, showing that most classes in current datasets are in fact static. However, in the bottom figure we observe that accuracy in the temporal classes is not as strongly correlated with the traditional accuracy, and in fact, it shows a "plateau" effect.

## 6. Effect of Training on Temporal Classes

We also explore the effect of training using temporal classes. In particular, we are interested in observing the behavior of the training in the absence of discriminative image information. In principle, this should force the network to learn stronger temporal features, which may help generalization. For the experiments we use the R(2+1)D architecture with RGB as input.

**Training on Temporal Classes avoids image bias and generalizes better.** In this experiment we measure the

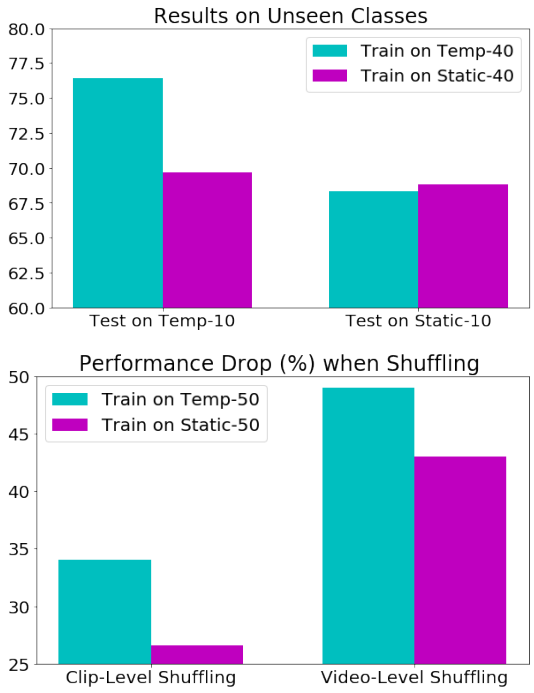



Figure 6. **Effect from training on temporal classes.** (a) Learning from temporal classes generalizes better to unseen classes than training on static classes. (b) Shuffling at test time has a much stronger effect on networks trained on temporal classes.

ability of a network to generalize to unseen classes when the network has been trained on temporal classes, and on static classes. We take each of the two datasets, which contain 50 classes, and we train on 40 of them, and leave the other 10 as unseen classes to test. After training a network on 40 classes, we freeze all weights except for the last layer, which we change to map to 10 classes, and we finetune it. Results are shown in Fig. 6 (a). We would expect that training on temporal classes would be better for testing on temporal classes, and that training on static would be better for testing on static. However, we find that testing on static shows on par performance for both training processes. We hypothesize that in the absence of discriminative image cues, the network cannot "cheat" and is forced to learn better temporal information, which is useful both for static and temporal classes. More importantly, we find that training on temporal performs much better testing on temporal, which shows the need of temporal data to learn good temporal features.

For a set of videos of unseen classes, we plot the activation maps [29] of the models trained on temporal and static classes, in Fig. 7. We observe that the activation maps of the model trained on temporal classes are much more meaningful than those from the model trained on static classes. The maps focus on the entire area where the action is happening, presumably because the network focuses on the motion, instead of the objects or scenes.

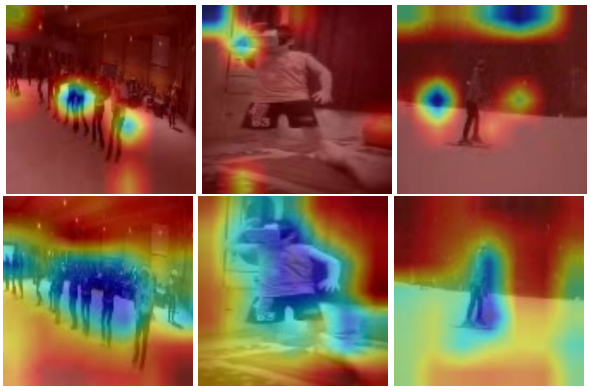


Figure 7. **Meaningful activation maps from models trained on static and temporal classes.** Each column shows frames of a particular video, corresponding to the classes "country line dancing", "robot dancing" and "skiing (not slalom)". The top row shows activation maps from a network trained on static classes, while the bottom row shows activation maps from a network trained on temporal classes. Both are tested on unseen classes. We observe that the activation maps from the network trained on temporal classes are more meaningful and centered around the action, even though these are unseen classes.

**Training on Temporal Classes produces features more sensitive to temporal changes.** Shuffling frames at test time has been noted to not harm performance dramatically [30]. This observation reveals that features are not very sensitive to changes in temporal structure. Does training on temporal classes produce more temporally-aware features? We measure the effect of shuffling frames on networks trained on static and temporal classes. Results are shown in Fig. 6 (b). We observe that the performance of a network trained on temporal classes drops much more than one trained on static classes. This is consistent both when shuffling at video and clip level. We argue that training on temporal classes leads to larger sensitivity to temporal ordering, and therefore stronger temporal features.

## 7. Conclusion

We have presented a methodology to discover the relevance of temporal information in action classes based on human perception. We use it to identify categories that contain temporal information and we called them temporal classes. We use temporal classes to benchmark various video models, revealing their abilities to model temporal information. We also use these classes for training and observe they boost accuracy on temporal categories, even on unseen classes. We hope that the proposed meta-dataset of temporal classes will help guide future research in temporal modeling for better video understanding.[2]

[2]Answers to first page riddle: "yawning".

## References


[1] *Kolmogorov–Smirnov Test*, pages 283–287. Springer New York, New York, NY, 2008.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[4] Anoop Cherian, Basura Fernando, Mehrtash Harandi, and Stephen Gould. Generalized rank pooling for activity recognition. *arXiv preprint arXiv*, 170402112, 2017.

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[6] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[7] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6202–6211, 2019.

[9] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):773–787, 2017.

[10] David F. Fouhey, Weicheng Kuo, Alexei A. Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018.

[11] Marianne Simmel Fritz Heider. An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243–259, 1944.

[12] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. *CoRR*, abs/1706.04261, 2017.

[13] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. *CoRR*, abs/1705.08421, 2017.

[14] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Limin Wang, and Shilei Wen. Stnet: Local and global spatial-temporal modeling for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8401–8408, 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video : Analyzing temporal information in video understanding models and datasets. In *ECCV*, 2018.

[17] YG Jiang, J Liu, A Roshan Zamir, G Toderici, I Laptev, M Shah, and R Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014.

[18] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211, 1973.

[19] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[20] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[21] Jan Koenderink, Whitman Richards, and Andrea "van Doorn". Space-time disarray and visual awareness. *i-Perception*, 3:159–165, 2012.

[22] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[23] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[24] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. *arXiv preprint arXiv:1811.08383*, 2018.

[25] Chenxu Luo and Alan L. Yuille. Grouped spatial-temporal aggregation for efficient action recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[26] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *ECCV (1)*, volume 9905 of *Lecture Notes in Computer Science*, pages 527–544. Springer, 2016.

[27] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown,

Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019.

[28] Jamie Ray, Heng Wang, Du Tran, Yufei Wang, Matt Feiszli, Lorenzo Torresani, and Manohar Paluri. Scenes-objects-actions: A multi-task, multi-label video dataset. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[30] Laura Sevilla-Lara, Yiyi Liao, Fatma Guney, Varun Jampani, Andreas Geiger, and Michael J. Black. On the integration of optical flow and action recognition. In *German Conference on Pattern Recognition (GCPR)*, Oct. 2018.

[31] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *CoRR*, abs/1604.01753, 2016.

[32] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576, 2014.

[33] K. Soomro, A. Roshan Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *CRCV-TR-12-01*, 2012.

[34] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

[35] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[36] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *CoRR*, abs/1711.11248, 2017.

[37] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[38] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018.

[39] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[40] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8668–8678, 2019.

[41] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. *European Conference on Computer Vision*, 2018.